Naval Health Research Center

(2)

TEST-RETEST RELIABILITY OF OXFORD MEDILOG 9000 SLEEP RECORDING AND SS-90-III SLEEP STAGE SCORING

D. G. McDONALD
L. IRWIN

REPORT NO. 89-6

Approved for public release: distribution unlimited.

NAVAL HEALTH RESEARCH CENTER
P.O. BOX 85122
SAN DIEGO, CALIFORNIA 92138

NAVAL MEDICAL RESEARCH AND DEVELOPMENT COMMAND
BETHESDA, MARYLAND

Test-Retest Reliability of Oxford Medilog 9000 Sleep Recording and SS-90-III Sleep Stage Scoring

David G. McDonald*
Lorene Irwin

*University of Missouri
Columbia, Missouri 65211

Naval Health Research Center
San Diego, California 92138-9174

Report No. 89-6, supported by the Naval Medical Research and Development Command, Bethesda, MD, Department of the Navy, under Work Unit MR4101-003-.6003. This research was conducted during the author's first appointment as an ASEE Summer Faculty Fellow, 1988. The opinions expressed in this paper are those of the authors and do not reflect official policy or position of the Navy, Department of Defense, not the U. S. Government. The authors wish to acknowledge the valuable assistance of Neal Morrison and Larry Matteson in the statistical analysis of the data, and Gary Welton who provided the method for the analysis of reliability of the data. Laverne Johnson, Steven A. Gomez, and Cheryl L. Spinweber also made important suggestions at various phases in this project.

SUMMARY

This study was undertaken to assess the test-retest reliability of the expanded Medilog SS-90-III Sleep Stager by comparing sleep stager scoring of the same records scored more than once. Subjects consisted of 19 normal sleepers who slept one night at home, following their normal routine. They ranged in age from 19.3 to 63.5, with mean = 31.1 and median = 24.7. There were 15 males and 4 females.

All sleep tapes were processed five consecutive times on the day of sleep scoring, termed runs 1-5. Dependent variables chosen for study consisted of: total sleep period, actual sleep time, wake after sleep onset (greater and less than 120 sec.), total movement time, total stage 1, total stage 2, total stage 3, total stage 4, total REM, percent of stage 1, percent of stage 2, percent of stage 3, percent of stage 4, percent of stage REM, start to sleep onset, sleep onset to REM, sleep onset to stage 2, sleep onset to stage 3, and sleep onset to stage 4.

Pearson correlations and alpha coefficients were calculated to evaluate the reliability of each measure over the five runs combined. Correlations between runs ranged between .878 and 1.00 for all measures, except sleep onset to REM through sleep onset to stage 4, where correlations ranged from -.003 to .991. Alpha coefficients ranged from .98 to 1.00 for all measures, including total sleep time, movement time, sleep onset, waking after sleep onset, and both absolute and percentage amounts of sleep stages 1-4 and REM, but not including the latency measures. Alpha coefficients for the latency measures were less reliable, although acceptable for REM latency, marginally so for stage 2 and 3 latencies, but not acceptable for stage 4 latency.

INTRODUCTION

Important technical advances in portable sleep recording and automated EEG sleep stage scoring methods have been introduced in recent years. These improved methods offer several economies for both subjects and investigators by virtue of the fact that sleep data are recorded on portable cassette tape EEG recorders. The recording electrodes can be applied almost any time day or night at the technologist's convenience, and the portability of the equipment means that subjects can choose to sleep at home, or in other, more familiar, environments. Further, the cassette tape can be scored in as little as 30-60 minutes and stored indefinitely as needed at minimal cost and space. Thus, this improved procedure has important advantages for (1) improved subject acceptance leading potentially to better sleep, (2) a more flexible workload for the technologist, and (3) quicker, yet reliable scoring of sleep records.

Evaluative reports in the literature have generally confirmed these advantages. Sewitch and Kupfer (1,2) have reported a project in which they compared the Oxford Medilog 9000 and Telediagnostic systems with laboratory recordings. They concluded that there were no differences in standard EEG sleep parameters recorded in either the home or laboratory. Hoelscher et al. (3) studied the usefulness of the Oxford Medilog system in evaluating a variety of sleep/wake disorders in a large clinical population and concluded that "the Oxford-Medilog 9000 can be used to evaluate a variety of sleep-related disorders, results in acceptable recordings in 90-97% of all studies, and is well-accepted by most patients" (p.607). However, both Sewitch and Kupfer and Hoelscher et al. noted difficulties in some cases, as well as remaining questions in scoring details.

More recently, Crawford (4) and Holler and Riemer (5) both compared visually-scored versus Medilog Sleep Stager-scored sleep records, in an effort to assess the accuracy and reliability of the automated method. Crawford found a range of 74 to 89% agreement (mean = 84% epoch by epoch) between manual and automated sleep scoring of 20 recordings. However, Holler and Riemer found consistent differences in (1) sleep onset time and (2) REM time, but not any other sleep measures, between manual and automated methods when applied to their sample of four sleep records. Neither Crawford nor Holler and Riemer compared automated sleep stager scoring of the same records scored more than once, i.e., correlation of the automated method with itself.

Further, while the reports of these investigators do indicate an acceptable level of reliability of automated sleep scoring by the Medilog Sleep Stager, it should be noted that the software supporting this system has been recently revised and expanded, thus requiring a revised and expanded evaluation of the scoring. The present study was therefore undertaken to evaluate the reliability of the expanded Medilog SS-90-III Sleep Stager by comparing sleep stager scoring of the same records scored more than once.

METHODS

Subjects

Subjects consisted of 21 healthy, normal sleepers recruited in the San Diego area. Four subject-night recordings were rejected because of electrode failure or inability of the subject to sleep more than four hours. Two of these were successfully rescheduled for a second night, and two were dropped from the study.

Therefore, the final N consisted of 19 subjects, ranging in age from 19.3 to 63.5 years, with mean = 31.1 and median = 24.7 years. There were 15 males and 4 females. Mean age of the females was not significantly different form the males (means of 32.7 and 30.9 years, respectively).

Procedures

Electrode attachment for most subjects was scheduled between 1400-1600 at the sleep laboratory, although in a minority of cases this was done at the subject's work location. Subjects then slept at home, following their normal routine, and returned to the laboratory about 0600-0700 for removal of the electrodes. Several subjects were given a bottle of acetone to remove the electrodes by themselves at home in the morning.

The standard sleep-monitoring montage consisted of the following: one channel of EEG (C4 to opposite mastoid), two channels of EOG (outer canthus to opposite mastoid), and one channel of EMG from the submental muscle of the chin. Clock time was used to define the beginning and end of the total sleep period (TSP).

Sleep scoring of all tapes was conducted 1-2 days after the recording night. All recommended procedures in the 1987 operator's manuals (Medilog 9000 Replay and Display System Operator's Instruction and Service Manual; Medilog SS-90-III Sleep Stager Operator's Manual) supplied by Oxford Medical Ltd. were followed closely. All tapes were processed on the Sleep Stager five consecutive times on the day of sleep scoring (i.e., five independent scoring runs always conducted on the same day), hereafter referred to as runs 1-5.

The Sleep Stager provides a printout with a hypnogram (see sample in Crawford (4)), plus the following measures: total sleep period (TSP), actual sleep time (AST), wake after sleep onset (WASO) > 120 sec., wake after sleep onset (WASO) < 120 sec., total movement time (TMT), total stage 1 (TOT1), total stage 2 (TOT2), total stage 3 (TOT3), total stage 4 (TOT4), total REM (TOTR), start to sleep onset (SSO), sleep onset to first REM (SOR), sleep onset to stage 2 (SO2), sleep onset to stage 3 (SO3), sleep onset to stage 4)SO4).

Of this list, all measures from AST to TOTR were provided in more than one unit of measurement or format, viz., by (1) number of epochs (30 secs.), (2) number of episodes (any occurrence lasting one epoch or more), (3) number of minutes, (4) percent of total sleep period, and (5) percent of actual sleep time (where appropriate). Thus, because of duplication, it was necessary to choose a smaller yet representative group of dependent variables for the present study. This group is listed in Table 1. All time measures are given in minutes. Latencies are calculated to the first four contiguous epochs (2 min.) of the stage. There is a partial duplication of measures only in that sleep stages 1 through REM are listed by both total time (TOT1-TOTR) and percent of actual time (PER1-PERR).

RESULTS

Means, standard deviations, and minimum and maximum values for all measures and all runs are summarized in Table 1. The minimum values listed are the lowest of the five minima of five runs, and the maximum values are the highest of the five maxima of five runs.

Pearson correlations between runs

Pearson correlation coefficients were calculated between consecutive sleep stage runs, which produced five correlation coefficients for each dependent variable. Runs were paired as follows: 1-2, 2-3, 3-4, 4-5, and 5-1. The range of these correlations between runs for each dependent variable is given in Table 1. It can be seen that the correlations between runs for the first 16 variables (TSP to SSO) were always between .878 and 1.00. Twenty-nine out of 32 correlations were ,90 or above, and the remaining three were above .87. For the four latency measures, SOR to SO4, the range of

Table 1: Summary of mean, standard deviation, minimum and aximum value, range of Pearson correlations between runs, and pha reliability coefficient for each of the dependent measures, l runs (1-5) combined.

MEASURE	MEAN	STD DEV	MIN	MAX	RANGE OF CO S BETW RUNS	ALPHA COEFF
TSP	476.11	81.28	330.0	653.5	.998-1.0	1.000
AST	400.44	71.32	265.0	542.0	.972-.99	.998
WASOG	89.83	71.89	0.0	265.0	.945-994	.994
WASOL	58.75	29.85	5.0	132.0	.899-.98	.989
TMT	2.75	4.93	0.0	18.0	.962-.99	.996
TOT1	119.90	119.74	2.0	422.0	.977-.99	.997
TOT2	261.46	132.40	0.0	460.0	.878-.99	.989
TOT3	41.41	35.12	0.0	131.0	.993-.99	.999
TOT4	49.14	60.99	0.0	215.0	.997-1.0	1.000
TOTR	328.91	155.24	59.0	677.0	.969-.99	.997
PER1	15.49	15.48	0.2	57.8	.976-.99	.997
PER2	32.30	15.17	0.0	54.8	.877-.99	.988
PER3	5.01	4.05	0.0	13.4	.990-.99	.999
PER4	5.79	6.90	0.0	19.9	.996-.99	1.000
PERR	41.39	17.86	6.4	82.9	.970-.99	.997
SSO	15.63	15.94	0.0	72.5	.944-.99	.994
SOR	56.64	58.57	0.0	432.5	.779-.97	.943
SO2	24.01	46.87	0.0	408.5	.097-.82	.645
SO3	41.49	78.82	0.0	438.0	.243-.83	.761
SO4	16.13	36.63	0.0	497.0	-.003-.99	.246

NOTE: N = 19; r of .43 = p<.05; r of .55 = p<.01. A reviations: TSP = total sleep period, AST = actual sleep time, WASO = waso > 120 sec., WASOL = waso < 120 sec, TMT = total movement time TOT1 = total stage 1, TOT2 = total stage 2, TOT3 = total stage 3, TC 4 = total stage 4, TOTR = total REM, PER1 = % stage 1, PER2 = perc t sgate 2, PER3 = percent stage 3, PER4 = percent stage 4, PERR = pe ent stage REM, SSO = start to sleep onset, SOR = sleep onset to REM, 2 = sleep onset to stage 2, SO3 = sleep onset to stage 3, SO4 = sleep nset to stage 4.

Pearson correlations were from -.003 to .991, and all o the three minimum correlations listed between runs for these measures fail to reach statistical significance at p<.05.

Alpha reliability coefficients

Alpha coefficients, using a method described by Cronbach (6), were calculated in order to evaluate the reliability of each measure over the five runs combined, based on the ratio of the variance of each individual measure to the variance of the composite of the five runs. The resulting alpha coefficients are given in Table 1. It can be seen that the alpha coefficients for the first 126 measures, TSP to SSO, are consistently .98 to 1.00, indicating very high reliability. For the latency measures (SOR to SO4), however, the alpha coefficients decline from .94 to .24, respectively, the latter failing to reach statistical significance in the case of SO4. While the alpha coefficients for SO4, SO2, and SO3 were significant at p<.01, one might still question whether the correlations of .64 and .76 for SO2 and SO3, respectively, are acceptable reliabilities for an automated scoring procedure.

DISCUSSION

The primary findings of this study may be summarized as follows: (1) Medilog SS-90-III Sleep Stager scoring of most sleep measures was highly reliable, with alpha coefficients ranging from .98 to 1.00 for measures including total sleep time, movement time, sleep onset, waking after sleep onset, and both absolute and percentage amounts of sleep stages 1-4 and REM; (2) scoring of latency measures was less reliable, although certainly acceptable for REM latency (SOR), and marginally acceptable for stage 2 and 3 latencies (SO2 and SO3), but not acceptable for an automated scoring of stage 4 latency (SO4).

While direct comparison with previous reports is difficult due to differences in procedures and methods of analysis, it would appear that our findings with the Medilog SS-90-III Sleep Stager represent a very considerable improvement over the results reported by both Crawford (4) and Holler and Riemer (5). Crawford found an average 74% agreement between manual and machine scoring across epochs for all measures, clearly suggesting a lower reliability, although she was not assessing retest reliability. Similarly, Holler and Riemer reported consistent error in measures of sleep onset and REM time, although neither of these variables appeared as problems in our data. While we did not compare manual and automated scoring provides a benchmark of retest reliability to which the manual method can only aspire.

Clear exceptions to this statement were the latency measures, SO2. SO3, and SO4 - especially the latter - where alpha coefficients were either marginally or unacceptably low. Reasons for this difficulty are undoubtedly that the stager defined latencies as the first four contiguous epochs of a given stage. Thus, for example, in measuring stage 4 onset, the scoring sequence 3444434443 would count, whereas 3444334443 would not, even though the overall accuracy of scoring was consistently high. Whenever the sleep stager missed the onset of a stage in this manner, and the subject did not show that stage again at all or until much later, then the latency value was greatly increased. Clearly, this effect would be larger in the case of a sleep stage that occurred most frequently, as was true of stage 4 sleep. Of the total of 285 reports of latency to onset sleep stages 2, 3, and 4, such shifts were observed in 10 cases, or 3.5%. Of this 10, three involved stage 2, three involved stage 3, and four involved stage 4.

REFERENCES

1. Sewitch DE, Kupfer DJ: Polysomnographic telemetry using Telediagnostic and Oxford Medilog systems. Sleep 1985; 8:288-293.

2. Sewitch DE, Kupfer DJ: A comparison of the Telediagnostic and Medilog systems for recording normal sleep in the home environment. Psychophysiol 1985; 22:718-726.

3. Hoelscher TJ, Erwin CW, Marsh GR, Webb MD, Radtke RA, Lininger A: Ambulatory sleep monitoring with the Oxford-Medilog 9000: Technical acceptability, patient acceptance, and clinical indications. Sleep 1987; 10:606-607.

4. Crawford C: Sleep recording in the home with automatic analysis of results. Eur Neurol 1986; 25:suppl 2, 30-35.

5. Holler L, Riemer, H: Comparison of visual analysis and automatic sleep stage scoring (Oxford Medilog 9000 system). Eur Neurol 1986; 25:suppl2, 36-45.

6. Cronbach L: Essentials of Psychological Testing, 4th Edition. New York: Harper and Row, 1984.

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a REPORT SECURITY CLASSIFICATION
Unclassified

1b RESTRICTIVE MARKINGS
N/A

2a. SECURITY CLASSIFICATION AUTHORITY
N/A

2b DECLASSIFICATION/DOWNGRADING SCHEDULE
N/A

3 DISTRIBUTION/AVAILABILITY OF REPORT
Approved for public release; distribution unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
NHRC Report No. 89-6

5 MONITORING ORGANIZATION REPORT NUMBER(S)

6a NAME OF PERFORMING ORGANIZATION
Naval Health Research Center

6b OFFICE SYMBOL *(If applicable)*
50

7a. NAME OF MONITORING ORGANIZATION
Commander Naval Medical Command

6c. ADDRESS *(City, State, and ZIP Code)*
P. O. Box 85122
San Diego, CA 92138-9174

7b. ADDRESS *(City, State, and ZIP Code)*
Department of the Navy
Washington DC 20371

8a NAME OF FUNDING/SPONSORING ORGANIZATION Naval Medical Research & Development Command

8b OFFICE SYMBOL *(If applicable)*

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER

8c. ADDRESS *(City, State, and ZIP Code)*
Naval Medical Command National Capital Region
Bethesda, MD 20814-5044

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO	WORK UNIT ACCESSION NO.
61155N	MR4101	003	6003

11 TITLE *(Include Security Classification)*
(U) TEST-RETEST RELIABILITY OF OXFORD MEDILOG 9000 SLEEP RECORDING AND SS-90-III SLEEP STAGE SCORING

12 PERSONAL AUTHOR(S)
DAVID G. MCDONALD AND LORENE IRWIN

13a TYPE OF REPORT
Interim

13b TIME COVERED
FROM ________ TO ______

14. DATE OF REPORT *(Year, Month, Day)*
1989March13

15 PAGE COUNT

16 SUPPLEMENTARY NOTATION
Prepared in cooperation with the University of Missouri-Columbia during an appointment of the senior author as an ASEE Summer Faculty Fellow, 1988.

17 COSATI CODES

FIELD	GROUP	SUB-GROUP

18 SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
Medilog system; Sleep scoring; EEG sleep; Portable recordings; Normal sleepers

19 ABSTRACT *(Continue on reverse if necessary and identify by block number)*

Sleep was recorded in 19 normal sleepers (19.3-63.5 years of age) one night at home, using the Medilog 9000 system to assess the reliability of the Medilog SS-90-III Sleep Stager by comparing sleep stager scoring of the same records scored five times. Primary results were: (1) Sleep Stager scoring of most sleep measures was highly reliable, with alpha coefficients ranging from .98 to 1.00 for total sleep time, movement time, sleep onset, waking after sleep onset, and both absolute and percentage amounts of sleep stages 1-4 and REM; (2) scoring of latency measures was less reliable, although certainly acceptable for REM latency and marginally acceptable for stage 2 and 3 latencies, but not acceptable for an automated scoring of stage 4 latency.

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT □ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a NAME OF RESPONSIBLE INDIVIDUAL
David McDonald, Ph.D./Lorene Irwin

22b TELEPHONE *(Include Area Code)*
(619) 532-6114

22c. OFFICE SYMBOL
50

DD FORM 1473, 84 MAR

83 APR edition may be used until exhausted.
All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE

*U.S. Government Printing Office: 1985-607-047